Corpora and ELT

Just a fling, or the real thing?

Costas Gabrielatos

Lancaster University, TESOL Greece

c.gabrielatos@lancaster.ac.uk

www.gabrielatos.com

INGED 2003

Baskent University, Ankara, Turkey

12th October 2003

- What is a corpus?
- How are corpora relevant to ELT?
- How can they be used?

What is a corpus?

Loose Definition

- Any body of text.

Common Definition

- A body of machine-readable text.

Strict Definition

- A finite collection of machine-readable text, sampled to be maximally representative of a language or variety.

Definitions from:
McEnery, T. & Wilson, A. (2001, 2nd ed.) Corpus Linguistics. Edinburgh University Press.

Types of corpora

- Reference - Monitor
- General - Specialised
- Whole text – Samples
- Spoken - Written
- Native speakers - Non-native speakers

How are corpora used?

Raw

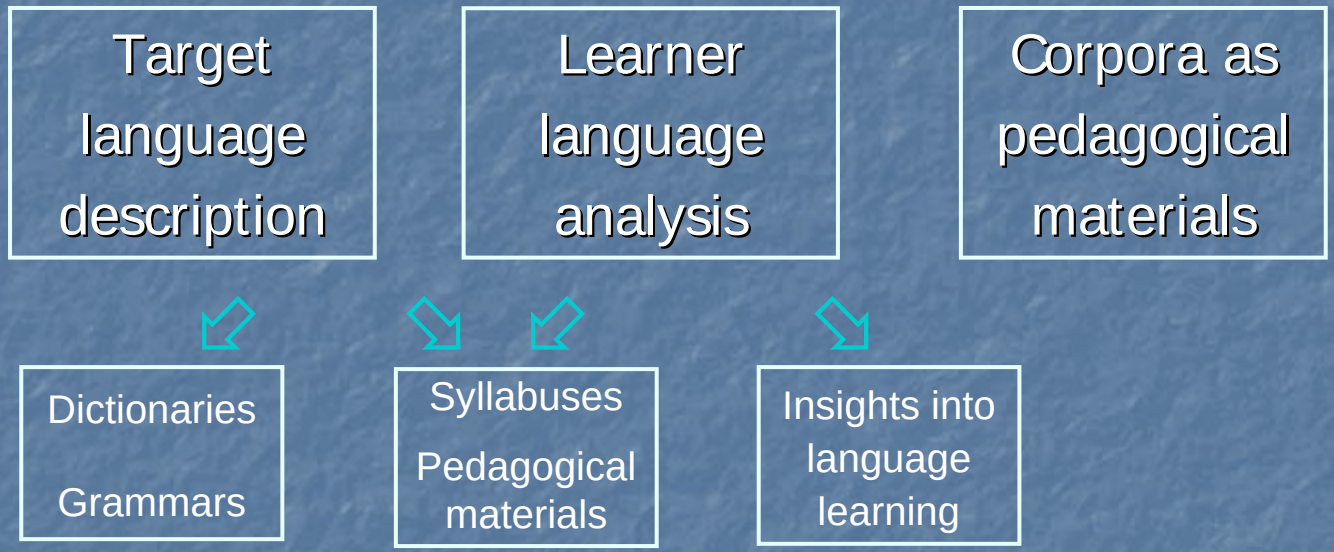
- Perhaps with sections/paragraphs indicated

Annotated

- Sections/paragraphs
- Part of speech
- Form (e.g. singular/plural, tense)
- Sense (e.g. lexis denoting belief, expectation)
- Pragmatic function (e.g. request, invitation)

0006526 910 The 93 AT 0006526 920 potential 03 [NN1/94] JJ/6 0006526 930 of 93 IO 0006526 940 computer 93 NN1 0006526 950 corpus 93 NN1 0006526 960 data-driven 33 [NN1/72] JJ/28 0006526 970 approaches 03 [NN2/71] VVZ/29 0006526 971 , 03 , 0006526 980 presenting 93 [VVG/99] NN1@/1 0006526 990 learners 03 NN2 0006526 000 with 93 IW 0006526 010 concordanced 06 [JJ@/100] VVD/0 VVN@/0 0006526 020 samples 93 [NN2/100] VVZ@/0 0006526 030 of 93 IO 0006526 040 language 03 NN1 0006526 050 for 93 [IF/100] CS%/0 0006526 060 analysis 03 NN1 0006526 061 , 03 , 0006526 070 comparison 93 NN1 0006526 080 and 93 CC 0006526 090 discussion 03 NN1 0006526 091 , 03 , 0006526 100 is 93 VBZ 0006526 110 also 93 RR 0006526 120 being 93 [VBG/100] NN1%/0 0006526 130 explored 93 [VVN/99] JJ@/1 VVD/0 0006526 131 .

How are corpora relevant to ELT?



Two additional benefits

- Computers are difficult learners ...
  - They can't handle ambiguity.
  - Everything has to be 'explained' clearly and in detail.
- Annotating corpora, i.e. teaching computers about language, can provide insights into ...
  - Language structure and use.
  - Language teaching.

Introspection

Intuitions

Data analysis

Quantification

Intuitions are not always dependable

"Question tags, along with bowler hats, mostly belong to 1960s BBC broadcasts."

Bradford, R. 2002. 'Grammar is by Statisticians, Language is by Humans.' IATEFL Issues 167 (p. 13).

"About every fourth question in conversation is a question tag."

Biber et al. 1999. Longman Grammar of Spoken and Written English. Longman. (p. 211). Based on the Longman Spoken and Written English Corpus (40 million words).

If the language information we give learners is based only on intuitions ...

If the examples/texts we select are chosen to reflect our intuitions ...

... then we may be presenting our personal informal observations about language, our preferred variety or idiosyncratic usage, as the only 'truth'.

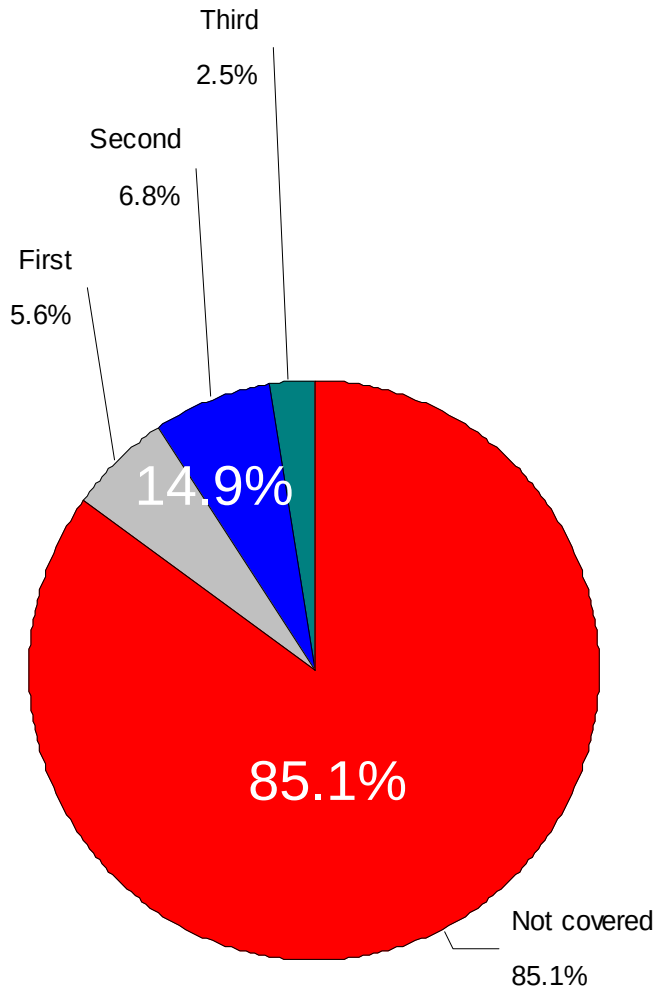
If-conditionals in a sample of 740 sentences from the written section of the British National Corpus

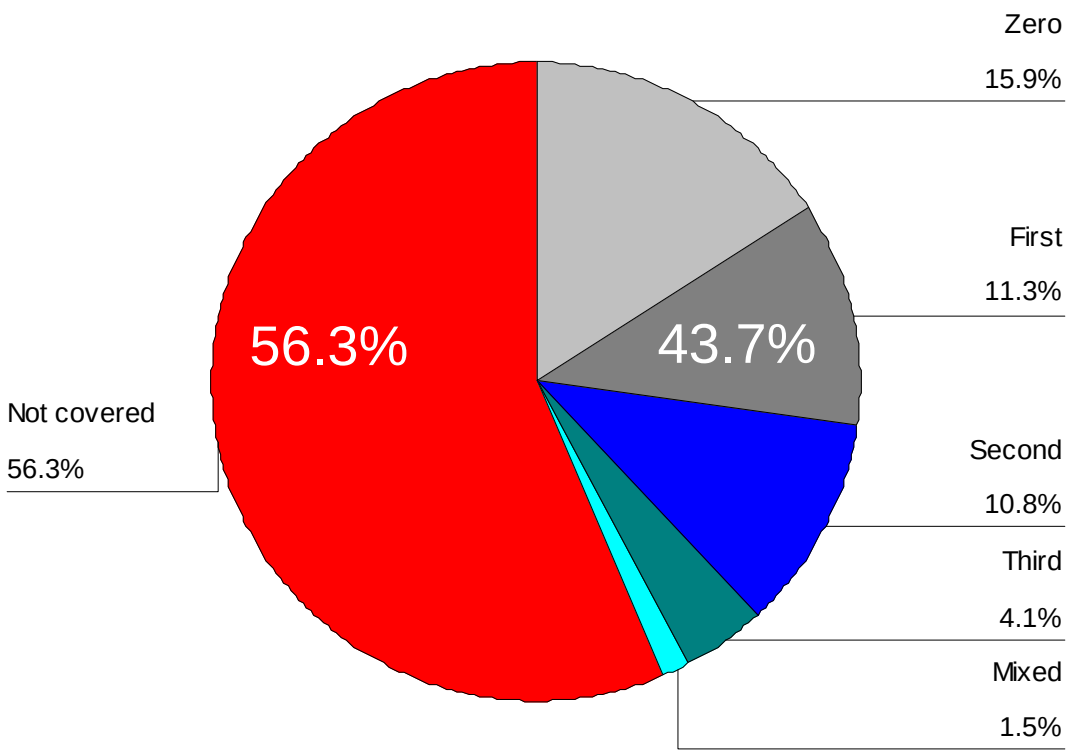
The three basic types

If + Present Simple	→	will + infinitive
If + Past Simple	→	would + infinitive
If + Past Perfect	→	would + perfect infinitive

The 5-types framework of if-conditionals, as presented in 20 coursebooks

(upper-intermediate to advanced)


Third
2.5%
Second
6.8%
First
5.6%
14.9%
85.1%
Not covered
85.1%



Uses of learner corpora

- Comparing NNS and NS use.
  - ➲ Over/under-use of features
  - ➲ Contexts of use
- Error analysis
  - ➲ Frequent / common errors
  - ➲ Patterns according to L1, age, medium, task type/context, genre

Intuitions as the target of learning

- Intuition, or a feel for the language, is what learners aim to develop.
- Native speakers have developed that feel through exposure to language in use and the recognition of patterns.
- Through this exposure native speakers have built the mental equivalent of a corpus.
- Intuitions can be seen as the results of the informal analysis of this mental corpus.

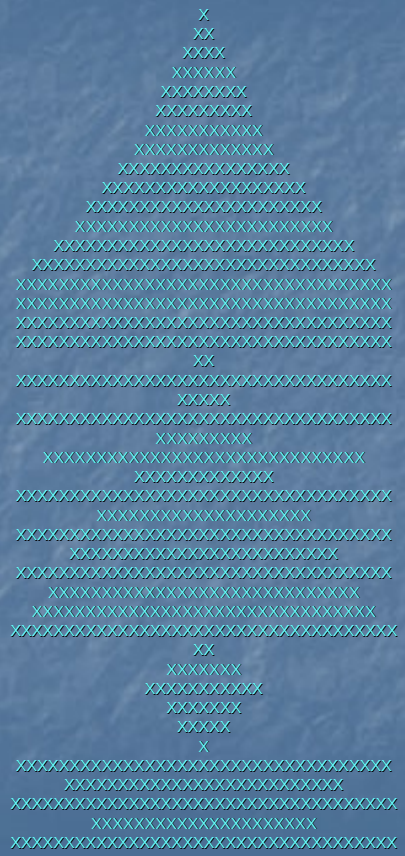
Visualising pattern recognition

- What are the parallels with teaching?
- What are the parallels with grammar rules and lexical explanations?

A small number of texts: a pattern

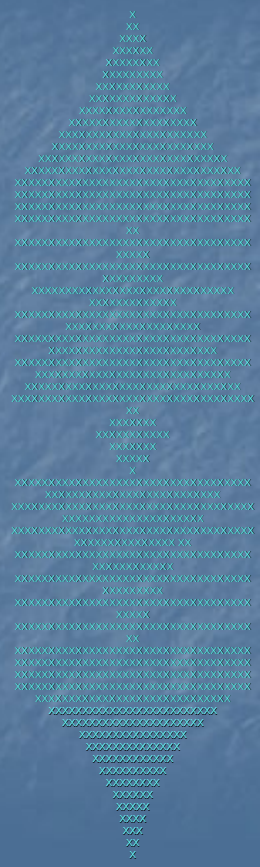
More texts: some irregularities

More texts: more irregularities ... or an emerging pattern?



More texts: a new pattern?

More texts will confirm



Looks like this is the pattern

More texts will confirm ...

How can corpora help with exposure?
The example of extensive reading

- EFL learners lack opportunities for rich language exposure and recognition of patterns.
- Extensive reading is seen as a good way to develop intuitions in the same way as native speakers.
- Representative corpora can offer 'condensed' exposure to such patterns.

Condensed language evidence

- One page contains 500 words on average.
- The BNC contains 100 million words.
- A six-year programme of five hours of teaching per week offers a total of 200 lessons.

- To get the same amount of language evidence on a specific lexical or grammatical point through reading the learner would need to ...
  - Examine 165 pages per lesson (intensive reading)
  - Read 90 pages per day (extensive reading), or 2-3 books per week.

Corpora in the classroom

Soft version

- Learners work with corpus-derived hard-copy materials

Hard version

- Learners work with corpora

Corpora in the classroom: prerequisites

- Teachers informed about corpora and trained in using corpus software.
- Corpus use incorporated in teacher preparation programmes.
- Learners trained in corpus use.
- Investment in corpora and software.

Corpora in the classroom:
Examples of the 'soft' version

- Collocations
- Lexical meaning and use
- Lexical inference

Your query "grief" returned 1396 matches, thinned with method *random selection* to 250 hits, sorted by *preceding* word at position 1 with tag-restriction *any adjective* (51 hits)

No	Filename	Solution 1 to 50 Page 1 / 2 Processed for lancaster03 at 81.135.71.178. 1		
1 (178)	H9Y 1540	childlike behaviour caused by the burden of **adult _AJ0-NN1**	**grief**	. On the night we were all united by one
2 (192)	HH0 1551	there quietly revising my priorities as far as **expending _AJ0-VVG**	**grief**	goes. After that my response seems mechanical.
3 (144)	G0T 1763	link between expressing anger and **expressing _AJ0-VVG**	**grief**	comes into play. The symbolism of weeping eczema
4 (39)	ARG 1404	, it seems allied with giving the **fundamental _AJ0**	**grief**	and despair room and expression, using them as a
5 (7)	AA8 61	the doctors "comments. Oh, **good _AJ0**	**grief**	.Their handwriting. "We can't read
6 (26)	AJT 104	shared the secrets of her lifestyle. **Good _AJ0**	**grief**	, we had a practising Buddhist in once to talk to
7 (58)	BNS 352	"**Good _AJ0**	**grief**	!" he exclaimed. "That seems a bit steep
8 (73)	CAF	years to come. **Good _AJ0**	**grief**	THE LAST OF ENGLAND Ted Walker Jonathan Cape,

No.	Word	Total No. in the whole BNC	As collocate	In No. of texts
There are 1241 different types in your collocation database for "sorrow". (Your query "sorrow" returned 548 matches in 312 different texts)				
1	joy	2956	23	19
2	anger	3712	20	20
3	pain	7332	16	16
4	death	20526	16	14
5	sorrow	570	14	5
6	people	123995	13	12
7	face	29030	7	7
8	happiness	1705	7	5

Collocations: exercise based on a single text

Read the text and ...

- Find (phrasal) verbs that combine with the word diet.
- Put them in some logical order.

  - prescribe
  - try
  - adopt
  - go on (2)
  - give up (2)
  - keep up

Collocations: exercise based on corpus samples

Examine the concordance and ...

- Find (phrasal) verbs that combine with the word diet.
- Group them in a logical way. Compare your groups.
- Does diet have the same meaning in every sentence?
- Which combinations seem to be more frequent with each meaning?
- Find combinations with similar/opposite meaning.

Hard version: getting more context

Once you have established a diet on which the child remains well, be careful not to allow too much of any one food.

Most children do grow out of their sensitivities gradually, and it is important not to keep them on a restricted diet any longer than necessary. Retest foods once or twice a year to see if they are still a problem. If the child has ever had a severe reaction, or suffers from asthma, then the retesting must be done very cautiously. Parents who have had one food-sensitive child will want to minimize their chances of having another, and some useful preventive measures are described in Chapter Thirteen.

Lexical meaning and use: homework

Examine the sentences with sorrow and grief

- What causes sorrow or grief?
- What other words/expressions with a similar or opposite meaning can you find?
- What verbs, adjectives and nouns seem to combine more often with the two words?
- Are there any frequent fixed expressions?

Teacher manipulation of corpus examples

- Writing / Speech
- Specific genre / text type
- Examples according to level and focus
- Amount of context

Potential problems: language description

- Corpus worship – discarding intuitions.
- Corpus studies depend on labelling and counting – which, in turn, depend on intuitions and theories.
- Generalising from non-representative corpora.
- Generalising from inappropriate corpora.

Potential problems: language teaching

- 'Doing corpora'.
- New prescription (e.g. frequency worship).
- Focus on lexis and grammar – neglecting language skills development.
- Focus on awareness – neglecting production.
- Low levels?
- Young learners?

Merits of corpora: language description 1

- Speed
- Accuracy
- Representativeness
- Quantification – frequencies
- The big picture – patterns
- Increasing availability – web access

Merits of corpora: language description 2

- Checking intuitions
- New insights - discovering patterns
- Idiosyncratic uses 'diluted' in corpus
- Enough data to examine idiosyncratic uses
- From prescription to description

Merits of corpora: learning / teaching

- From rule to patterns
- Actual rather than made-up examples
- A wealth of examples in different contexts
- Amount of context according to needs
- New lease of life for the language lab
- Data-driven / discovery learning
- Learner-centred methodology: the learner as language researcher/detective

Corpus use has shown potential to ...

- Provide increasingly clear and accurate descriptions of native and learner language.
- Empower NNS teachers and researchers.
- Enhance data-driven / discovery learning
- Reinforce learner-centred methodologies